The Journal of Open Source Software

{fairmetrics}: An R package for group fairness evaluation

Benjamin Smith[1], **Jianhui Gao**[1], **and Jessica Gronsbell**[1]

1 Department of Statistical Science, University of Toronto

DOI:

Software

- Review
- Repository
- Archive

Submitted:
Published:

Summary

Fairness is a growing area of machine learning (ML) that focuses on ensuring that models do not produce systematically biased outcomes across groups defined by protected attributes, such as race, gender, or age. The {fairmetrics} R package provides a user-friendly framework for rigorously evaluating group-based fairness criteria, including independence (e.g., statistical parity), separation (e.g., equalized odds), and sufficiency (e.g., predictive parity). The package provides both point and interval estimates for a variety of commonly used criteria. {fairmetrics} also includes an example dataset derived from the Medical Information Mart for Intensive Care, version II (MIMIC-II) database (Goldberger et al. 2000; J. Raffa 2016) to demonstrate its use.

Statement of Need

ML models are increasingly used in high-stakes domains such as criminal justice, healthcare, finance, employment, and education Mattu (2016). Existing fairness evaluation software report point estimates and/or visualizations, without any measures of uncertainty. This limits users' ability to determine whether observed disparities are statistically significant. {fairmetrics} addresses this limitation by including confidence intervals for both difference and ratio based fairness metrics to enable more robust and statistically grounded fairness assessments.

Fairness Criteria

{fairmetrics} is designed to evaluate fairness of binary classification models across binary protected attributes. The package supports the evaluation of metrics belonging to three major group fairness criteria:

- **Independence:** Statistical Parity (compares the overall rate of positive predictions between groups), Conditional Statistical Parity (compares the rate of positive predictions between groups within a specific subgroup).
- **Separation:** Equal Opportunity (compares false negative rates between groups), Predictive Equality (compares false positive rates between groups), Balance for Positive Class (compares the average predicted probabilities among individuals whose true outcome is positive across groups), and Balance for Negative Class (compares the average predicted probabilities among individuals whose true outcome is negative across groups).

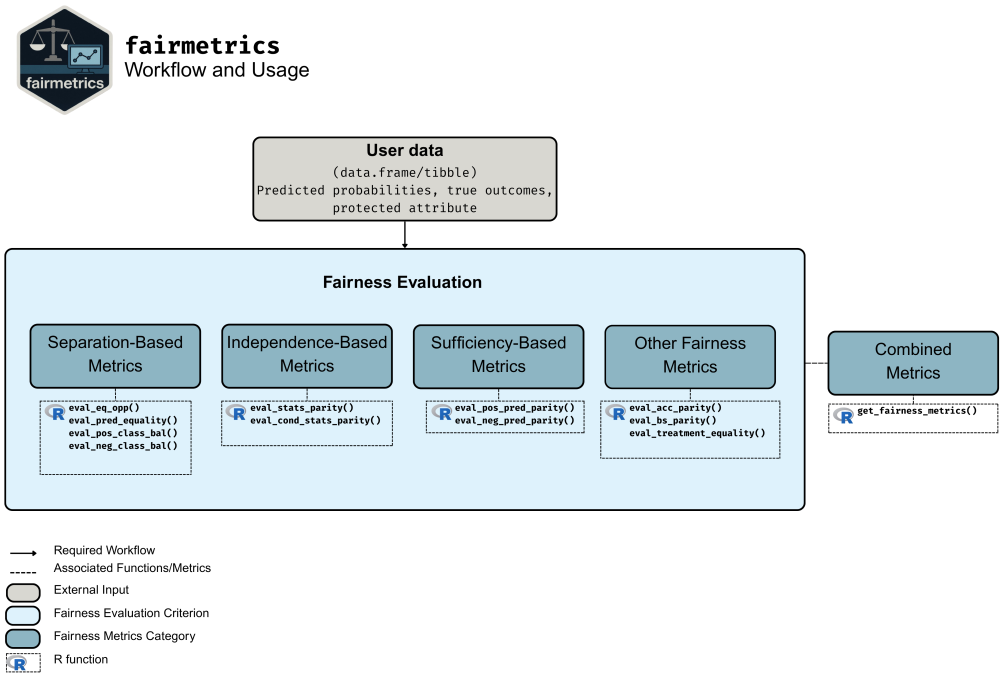

Figure 1: Workflow for using {fairmetrics} to evaluate model fairness across multiple criteria.

- **Sufficiency:** Positive Predictive Parity (compares the positive predictive values across groups), Negative Predictive Parity (compares the negative predictive values across groups).

The package also includes additional metrics, such as the Brier Score Parity (compares the Brier score across groups), Accuracy Parity (compares the overall accuracy across groups), and Treatment Equality (compares the ratio of false negatives to false positives across groups).

Evaluating Fairness Criteria

The input required to evaluate model fairness with the {fairmetrics} package is a data frame or tibble containing the model's predicted probabilities, the true outcomes, and the protected attribute. Figure 1 shows the workflow for using {fairmetrics}.

A simple example of how to use the {fairmetrics} package is illustrated below. The example makes use of the `mimic_preprocessed` dataset, a pre-processed version of the Indwelling Arterial Catheter (IAC) Clinical dataset, from the MIMIC-II clinical database (Goldberger et al. 2000; J. Raffa 2016; J. D. Raffa et al. 2016).

While the choice of fairness metric used is context dependent, we show all criteria available with the `get_fairness_metrics()` function for illustrative purposes. In this example, we evaluate the model's fairness with respect to the protected attribute `gender`. For conditional statistical parity, we condition on age greater than 60 years. The model is trained on a subset of the data and the predictions are made and evaluated on a test set. A statistically significant difference across groups at a given level of significance is indicated when the confidence interval for a difference-based metric does not include zero or when the interval for a ratio-based metric does not include one.

```r
# Train a classification model (e.g., random forest).
# Add the vector of predicted probabilities to the test data
# to evaluate fairness.
library(fairmetrics)
# Setting alpha=0.05 for 95% confidence intervals
get_fairness_metrics(
 data = test_data,
 outcome = "day_28_flg",
 group = "gender",
 group2 = "age",
 condition = ">=60",
 probs = "pred",
 cutoff = 0.41,
 alpha = 0.05
)

$performance
                              Metric GroupFemale GroupMale
1           Positive Prediction Rate        0.18      0.09
2           Positive Prediction Rate        0.35      0.24
3                False Negative Rate        0.36      0.59
4                False Positive Rate        0.08      0.03
5                Avg. Predicted Prob.       0.47      0.37
6                Avg. Predicted Prob.       0.16      0.11
7          Positive Predictive Value        0.62      0.67
8          Negative Predictive Value        0.93      0.91
9                        Brier Score        0.09      0.08
10                          Accuracy        0.87      0.89
11 (False Negative)/(False Positive) Ratio  0.93      2.89

$fairness
                          Metric Difference    95% Diff CI Ratio 95% Ratio CI
1             Statistical Parity       0.08   [0.04, 0.12]  2.00 [1.41, 2.83]
2 Conditional Statistical Parity       0.10   [0.02, 0.18]  1.43 [1.05, 1.96]
3              Equal Opportunity      -0.21 [-0.36, -0.06]  0.64  [0.45, 0.9]
4           Predictive Equality        0.04   [0.01, 0.07]  2.33  [1.13, 4.8]
5     Balance for Positive Class       0.09   [0.04, 0.14]  1.24 [1.09, 1.42]
6     Balance for Negative Class       0.05   [0.03, 0.07]  1.50 [1.29, 1.75]
7      Positive Predictive Parity     -0.06  [-0.23, 0.11]  0.92 [0.71, 1.18]
8      Negative Predictive Parity      0.01  [-0.15, 0.17]  1.01 [0.79, 1.29]
9             Brier Score Parity       0.01  [-0.01, 0.03]  1.12 [0.89, 1.43]
10        Overall Accuracy Parity     -0.01  [-0.05, 0.03]  0.99 [0.95, 1.03]
11            Treatment Equality      -2.28  [-4.72, 0.16]  0.33 [0.15, 0.73]
```

Users can also compute individual metrics using functions like `eval_eq_opp()` to test specific fairness conditions. Full usage examples are provided in the package documentation.

Related Work

Other R packages similar to {fairmetrics} include {fairness} (Kozodoi and V. Varga 2021), {fairmodels} (Wiśniewski and Biecek 2022) and {mlr3fairness} (Pfisterer, Siyi, and Lang 2024). {fairmetrics} differs from these packages in two ways. The first difference is that {fairmetrics} calculates ratio and difference-based group fairness metrics and their corresponding confidence intervals, allowing for more meaningful inferences about the fairness criteria. The second difference is that {fairmetrics} does not possess any external dependencies and has a lower memory footprint. Table 1 shows the comparison of memory used and dependencies required when loading each library.

Package	Memory (MB)	Dependencies
fairmodels	17.02	29
fairness	117.61	141
mlr3fairness	58.11	45
fairmetrics	0.05	0

Table 1: Memory usage (in MB) and dependencies of fairmetrics vs similar packages.

For Python users, the {fairlearn} library (Weerts et al. 2023) provides additional fairness metrics and algorithms. The {fairmetrics} package is designed for seamless integration with R workflows, making it a more convenient choice for R users.

Licensing and Availability

The {fairmetrics} package is under the MIT license. It is available on CRAN and can be installed by using `install.packages("fairmetrics")`. Full documentation and its examples are available at: https://jianhuig.github.io/fairmetrics/articles/fairmetrics.html. Source code and issue tracking are hosted on GitHub: https://github.com/jianhuig/fairmetrics/.

References

Anderson, Joshua W, and Shyam Visweswaran. 2024. "Algorithmic Individual Fairness and Healthcare: A Scoping Review." *JAMIA Open* 8 (1). https://doi.org/10.1093/jamiaopen/ooae149.

Awasthi, Pranjal, Corinna Cortes, Yishay Mansour, and Mehryar Mohri. 2020. "Beyond Individual and Group Fairness." *arXiv.org*. https://arxiv.org/abs/2008.09490.

Barocas, Solon, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. Cambridge, Massachusetts: The MIT Press.

Berk, Richard, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2018. "Fairness in Criminal Justice Risk Assessments: The State of the Art." *Sociological Methods & Research* 50 (1): 3–44. https://doi.org/10.1177/0049124118782533.

Castelnovo, Alessandro, Riccardo Crupi, Greta Greco, Daniele Regoli, Ilaria Giuseppina Penco, and Andrea Claudio Cosentini. 2022. "A Clarification of the Nuances in the Fairness Metrics Landscape." *Scientific Reports* 12 (1). https://doi.org/10.1038/s41598-022-07939-1.

Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. "Fairness Through Awareness." In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–26. ITCS '12. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2090236.2090255.

Fazelpour, Sina, and David Danks. 2021. "Algorithmic Bias: Senses, Sources, Solutions." *Philosophy Compass* 16 (8). https://doi.org/10.1111/phc3.12760.

Gao, Jianhui, Benson Chou, Zachary R. McCaw, Hilary Thurston, Paul Varghese, Chuan Hong, and Jessica Gronsbell. 2024. "What Is Fair? Defining Fairness in Machine Learning for Health." *arXiv.org*. https://arxiv.org/abs/2406.09307.

Goldberger, Ary L., Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. 2000. "PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals." *Circulation [Online]* 101 (23): e215–20. https://doi.org/10.1161/01.CIR.101.23.e215.

Grolemund, Hadley Wickham, and Garrett. n.d. "10 Tibbles | r for Data Science." https://r4ds.had.co.nz/tibbles.html.

Grote, Thomas, and Geoff Keeling. 2022. "Enabling Fairness in Healthcare Through Machine Learning." *Ethics and Information Technology* 24 (3): 39. https://doi.org/10.1007/s10676-022-09658-7.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY: Springer New York. https://doi.org/10.1007/978-0-387-84858-7.

Initiative, Open Source. n.d. "The MIT License." *Open Source Initiative*. https://opensource.org/license/mit.

Kozodoi, Nikita, and Tibor V. Varga. 2021. *Fairness: Algorithmic Fairness Metrics*. https://CRAN.R-project.org/package=fairness.

Lum, Kristian, Yunfeng Zhang, and Amanda Bower. 2022. "De-Biasing 'Bias' Measurement." In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 379–89. FAccT '22. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3531146.3533105.

Makhlouf, Karima, Sami Zhioua, and Catuscia Palamidessi. 2024. "When Causality Meets Fairness: A Survey." *Journal of Logical and Algebraic Methods in Programming* 141 (June): 101000. https://doi.org/10.1016/j.jlamp.2024.101000.

Mattu, Lauren Kirchner, Jeff Larson. 2016. "Machine Bias." *ProPublica*. https://www.propublica.o bias-risk-assessments-in-criminal-sentencing.

Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021a. "A Survey on Bias and Fairness in Machine Learning." *ACM Comput. Surv.* 54 (6). https://doi.org/10.1145/3457607.

———. 2021b. "A Survey on Bias and Fairness in Machine Learning." *ACM Computing Surveys* 54 (6): 115:1–35. https://doi.org/10.1145/3457607.

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations." *Science* 366 (6464): 447–53.

Pfisterer, Florian, Wei Siyi, and Michel Lang. 2024. *Mlr3fairness: Fairness Auditing and Debiasing for 'Mlr3'*. https://mlr3fairness.mlr-org.com.

Plecko, Drago, and Elias Bareinboim. 2022. "Causal Fairness Analysis." *arXiv.org*. https://arxiv.org/abs/2207.11385.

Raffa, Jesse. 2016. "Clinical Data from the MIMIC-II Database for a Case Study on Indwelling Arterial Catheters (Version 1.0)." https://doi.org/10.13026/C2NC7F. https://doi.org/10.13026/C2NC7F.

Raffa, Jesse D., Mohammad Ghassemi, Tristan Naumann, Mengling Feng, and Daniel J. Hsu. 2016. "Data Analysis." In *Secondary Analysis of Electronic Health Records*, 109–22. Springer, Cham. https://doi.org/10.1007/978-3-319-43742-2_9.

Rajpurkar, Pranav, Emma Chen, Oishi Banerjee, and Eric J. Topol. 2022. "AI in Health and Medicine." *Nature Medicine* 28 (1): 31–38. https://doi.org/10.1038/s41591-021-01614-0.

Ueda, Daiju, Taichi Kakinuma, Shohei Fujita, Koji Kamagata, Yasutaka Fushimi, Rintaro Ito, Yusuke Matsui, et al. 2023. "Fairness of Artificial Intelligence in Healthcare: Review and Recommendations." *Japanese Journal of Radiology* 42 (1): 3–15. https://doi.org/10.1007/s11604-023-01474-3.

Weerts, Hilde, Miroslav Dudík, Richard Edgar, Adrin Jalali, Roman Lutz, and Michael Madaio. 2023. "FairLearn: Assessing and Improving Fairness of AI Systems." *arXiv.org*. https://arxiv.org/abs/2303.16626.

Wickham, Hadley, Jim Hester, Winston Chang, and Jennifer Bryan. 2022. *Devtools:*

Tools to Make Developing r Packages Easier. https://CRAN.R-project.org/package=devtools.

Wiśniewski, Jakub, and Przemysław Biecek. 2022. "Fairmodels: A Flexible Tool for Bias Detection, Visualization, and Mitigation in Binary Classification Models." *The R Journal* 14 (1): 227–43. https://doi.org/10.32614/RJ-2022-019.

Yfantidou, Sofia, Marios Constantinides, Dimitris Spathis, Athena Vakali, Daniele Quercia, and Fahim Kawsar. 2023. "Beyond Accuracy: A Critical Review of Fairness in Machine Learning for Mobile and Wearable Computing." *arXiv.org*. https://arxiv.org/abs/2303.15585.